GLAMM: Genome-Linked Application for Metabolic Maps

John T. Bates
Physical Biosciences Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Dylan Chivian
Physical Biosciences Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Adam P. Arkin
Department of Bioengineering
University of California, Berkeley
and
Physical Biosciences Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

This work was supported by the Director, Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

GLAMM: Genome-Linked Application for Metabolic Maps

John T. Bates[1,2], Dylan Chivian[1,2,*], Adam P. Arkin[1,2,3,*]

[1]Technologies Division, DOE Joint BioEnergy Institute, Emeryville, CA 94608 USA;
[2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720 USA;
[3]Department of Bioengineering, University of California, Berkeley, CA, 94720 USA

* Please address correspondence to DCChivian@lbl.gov and APArkin@lbl.gov.
Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., MS-978, Berkeley, CA 94720 USA

Abstract

The Genome-Linked Application for Metabolic Maps (GLAMM) is a unified web interface for visualizing metabolic networks, reconstructing metabolic networks from annotated genome data, visualizing experimental data in the context of metabolic networks, and investigating the construction of novel, transgenic pathways. This simple, user-friendly interface is tightly integrated with the comparative genomics tools of MicrobesOnline (1). GLAMM is available for free to the scientific community at glamm.lbl.gov.

Introduction

As the volume of genomic, experimental, and metabolic network data increases, so has the need for clean, unobtrusive methods for visualizing and contextualizing these data. With this in mind, we have developed the Genome-Linked Application for Metabolic Maps (GLAMM). GLAMM provides a unified web interface for visualizing metabolic networks, reconstructing metabolic networks from annotated genome data or custom user-defined networks, visualizing experimental data in the context of metabolic networks, and investigating the construction of novel, transgenic pathways.

Other web resources (2-7) such as the KEGG Atlas, iPath, Pathway Projector, MetaCyc, and Reactome offer similar, web-based mapping style interfaces, but GLAMM also incorporates an interface for biological retrosynthesis (8-10), visualization of thousands of publicly accessible experimental or other user-defined data in the context of metabolic pathways, and integration with MicrobesOnline (1). This integration provides GLAMM users access to MicrobesOnline's powerful comparative phylo-genomic and functional genomic tools and a database of nearly 2000 prokaryotic and fungal genomes, allowing rapid analysis of genome context, regulon discovery, and so on.

GLAMM was developed using the Google^{TM} Web Toolkit (GWT, http://code.google.com/webtoolkit/) for the client UI and server implementation. The underlying maps are scalable vector graphics (SVG) documents rendered in real time on the client side in a GWT widget, with UI components and event handling provided by the GWT. Both of these technologies have the advantage of consistent cross-browser support, as well as a highly optimized execution path, with JavaScript and SVG rendered by the browser's own internal implementations. As such, GLAMM will only work with browsers that support both JavaScript and SVG (e.g. Firefox, Chrome, and Safari). This implementation performs well for thousands of on-screen elements on a typical personal computer.

In addition to a client-side interface, we have implemented a server that is integrated with MicrobesOnline. The GLAMM server communicates with the client via highly modularized and separable XML. The client can request any combination of pathways, reactions, genes, and compounds. It also can request functional data, currently gene expression data, but data associated with reactions (e.g. flux) and metabolites (e.g. concentrations) will be supported in the near future. We chose to create a new, lightweight XML format that only included the features needed by the interface rather than employ an existing format such as SBML (11) or BioPAX (12) in order to minimize the data necessary to transfer and because we needed to add support for features not already captured by SBML or BioPAX. We expect to support export to BioPAX and perhaps SBML in the future.

Underlying Metabolic Network

We have developed a method for aggregating and normalizing compound, reaction, and pathway data from several different metabolic databases. We chose to first focus our attention on combining KEGG (13), MetaCyc (5), and the compound and reaction databases provided for the E. coli iJR904 model (7). We also included reconciliation of metabolites with PubChem (14) and ChEBI (15). The database aggregation and normalization code is general enough to accommodate information from any similar database with the addition of a compatible parser with an eye toward inclusion of custom pathways, such as those found in organisms of interest to bioremediation and biofuel production.

Compounds and reactions were extracted from flat-file representations of the databases and converted to a normal form. For compounds, this normal form includes information such as common name, mass, formula, SMILES representation (16), InChI representation (17), compound name synonyms, and external references to the compound in other databases. Similarly, for reactions, the normal form includes a normalized form of the balanced reaction equation, a human-readable reaction definition, external references to this reaction in other databases, E.C. numbers (18) (if applicable), KEGG RPAIR role information for the reactants and products, and the KEGG pathway to which this reaction belongs. The normalized format is flexible enough to be expanded as custom reactions are introduced.

Compounds present in multiple databases are resolved into single entries by comparing the external reference IDs (e.g. PubChem) and merging normalized entries if a match is found. For consistency, KEGG common names, masses, and formulae take precedence over those from other databases. We are continuing to investigate schemes for normalizing reactions, a more complicated endeavor as a consequence of the numerous similar but non-identical names given to reactants, products, and secondary metabolites and which are included in the definition of each reaction (e.g. the inclusion of chirality information, different protonation states, polymers, etc.).

The data aggregation and normalization code is written entirely in object-oriented Perl and therefore can be run on almost any platform. This will no doubt change, as we intend to develop a fully automatic update and reconciliation mechanism. While the individual databases we have incorporated are curated, there remain some reactions that do not always account for mass balance or possess other eccentricities. Regrettably, it is beyond the scope of this project to rectify those issues, but we will update our imported network as improved data becomes available.

Automated Metabolic Reconstruction

GLAMM uses the gene annotations in MicrobesOnline to automatically reconstruct the metabolic networks of almost 2000 organisms. It combines MicrobesOnline's E.C. assignments (derived from hits to TIGRFAMs (19), KEGG annotations, and orthologs from reference genomes) with the E.C. number to reaction ID mappings from the public databases aggregated in GLAMM. Taken together, these mappings loosely determine the set of reactions available for a given organism. We recognize that automated E.C. assignments based solely on homology to a gene family are limited and by no means comparable to that of dedicated reconstruction pipelines such as ModelSEED (20) or manually-curated reconstructions (7). GLAMM therefore supports custom, user-uploaded reconstructions (see below) and will support reconstructions from other databases in the future.

When the user selects a host organism, GLAMM prunes the set of reaction edges in the global map to only include those reactions available to that organism (Figure 1). The remaining reaction edges on the displayed map are grayed out. Based on the connectivity information supplied with the map, GLAMM also prunes compound nodes that have no reactions associated with them. This not only yields the metabolic reaction network, but also the set of all compounds endogenous to the host, within the constraints of the displayed map which is, by necessity, a subset of the actual metabolic network of known chemical transformations.

There are obvious limitations to this technique, including the incompleteness of E.C. assignments for genes and that E.C. numbers often specify a broad class of reactions and therefore may not be substrate-specific. We aim to overcome these limitations in the future by augmenting the MicrobesOnline database with direct gene to reaction mappings (e.g. using KEGG orthologs.)

Custom Metabolic Reconstruction

GLAMM also provides a mechanism for uploading custom metabolic networks. Initially, this is in the form of tab-delimited files containing gene ID to E.C. number or gene ID to reaction ID mappings. Eventually we aim to support SBML and BioPAX specified pathways directly. The default metabolic reconstruction for any organism in MicrobesOnline may be downloaded, modified, and re-uploaded.

GLAMM Feature Highlights

Metabolites and Metabolic Pathways

The current GLAMM global view presents the KEGG Atlas map, but can be updated with any metabolic map using a standard format that we have designed. The resultant visualized map is pannable and zoomable as typical of web mapping applications. Compounds are represented as nodes on the map. Reactions, along with their corresponding genes in the organism-specific metabolic network reconstructions, are represented as edges. Clicking on the nodes presents a popup window (Figure 1) containing the compound's name, its formula, its mass, and a structural diagram, if available. Similarly, clicking on the edges of the map presents a popup window containing the reaction's human-readable definitions, its E.C. numbers, and the number of genes corresponding to those E.C. numbers in the target organism. The global map also contains textual labels for the various subpathways, and clicking on those labels presents popup windows containing schematic representations of the more detailed KEGG pathway maps. All popups include links back to the corresponding pathways, genes, or metabolites in MicrobesOnline.

Route Finding and Retrosynthesis

For convenience, we have included a search dialog box that re-centers the map around any compound, reaction, or gene name specified by the user. Additionally, the global view will allow the user to "get directions" in finding optimal pathways between a starting metabolite and a desired target metabolite (Figure 2). In the event of ambiguous compound search results, often due to the presence of multiple isomers on the map (e.g. glucose may specify alpha-D-glucose or beta-D-glucose,) a disambiguation popup will appear, allowing the user to specify the desired compound. Suggested pathways may offer routes for retrosynthesis and traverse all annotated organisms or otherwise conceivable reaction steps using a variety of appropriate pathway/gene set cost functions, returning the necessary genes to add to the host in order to complete the pathway from the chassis network to the target molecule. The routes are overlaid on top of the main map view, and all non-participating reactions are grayed out. If a host organism is selected, the E.C. number links to MicrobesOnline for candidate genes and retrosynthetic pathways are enabled in order to facilitate further examination with its powerful comparative systems biology tools, including gene trees, genome context and operon predictions, functional residue alignments, basic structural models, and functional expression data. These tools are provided with the intent of developing a mutually consistent set of genes for introducing the pathway into the host organism.

Experimental Data Visualization

Additionally, the global view can be used to visualize any data as an “overlay”, including *omics data such gene expression, protein levels, flux, source organism for a given reaction in a synthetic network, kinetic and thermodynamic parameters, optimal paths between metabolites, and so on (Figure 3). For example, *omics data will permit the user to analyze the global behavior of the network when challenged by stressful conditions or particular nutrient levels and to identify key pathways that are either directly involved in target molecule synthesis or may otherwise impact metabolic engineering.

Custom Data Overlay

In addition to public experimental data available on MicrobesOnline, the user may upload tab-delimited files with a list of genes and numerical data values for those genes. Similar to the downloadable metabolic reconstructions, one may also download experimental data sets that contain gene names consistent with metabolic reconstructions, to which new data values may be applied.

Future Directions

GLAMM will continue to be developed to support additional data types and custom display of data associated with reactions and metabolites. Additional bounds on retrosynthesis pathways, as well as longer pathways will be implemented to permit the user to require routes that pass through or avoid user-defined intermediates, that maximize or minimize use of particular cofactors, that maximize predicted flux, and so on. Source code will be made available freely for academic research.

Acknowledgements

This work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 with the E.O. Lawrence Berkeley National Laboratory (LBNL). Additional support was received from the "Knowledgebase R&D" project, sponsored by the Office of Biological and Environmental Research in the DOE Office of Science with American Recovery and Reinvestment Act funding to Oak Ridge National Laboratory (ORNL) and performed at LBNL. ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

The authors would also like to thank Thanya Suwansawad for the design of the GLAMM logo.

References

1. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S., Dubchak,I.L., Alm,E.J., and Arkin,A.P. (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*. **38**, D396-400. doi:10.1093/nar/gkp919
2. Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S., and Kanehisa,M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*. **36**, W423-6. doi:10.1093/nar/gkn282
3. Letunic,I., Yamada,T., Kanehisa,M., and Bork,P. (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*. **33**, 101-3. doi:10.1016/j.tibs.2008.01.001
4. Kono,N., Arakawa,K., Ogawa,R., Kido,N., Oshita,K., Ikegami,K., Tamaki,S., and Tomita,M. (2009) Pathway Projector: Web-Based Zoomable Pathway Browser Using KEGG Atlas and Google Maps API. *PLoS One* **4**, e7710. PMID: 19907644
5. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A., Paley,S., Popescu,L., Pujar,A., Shearer,A.G., Zhang,P., and Karp,P.D. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. **38**, D473-9. doi:10.1002/0471250953.bi0117s20
6. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B., Jupe,S., Kalatskaya,I., Mahajan,S., May,B., Ndegwa,N., Schmidt,E., Shamovsky,V., Yung,C., Birney,E., Hermjakob,H., D'Eustachio,P., and Stein,L. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. **39**, D691-7. doi:10.1093/nar/gkq1018
7. Schellenberger,J., Park,J.O., Conrad,T.C., and Palsson,B.Ø. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions, *BMC Bioinformatics* **11**, 213. doi:10.1186/1471-2105-11-213
8. Prather,K.L.J. and Martin,C.H. (2008) De novo biosynthetic pathways: rational design of microbial chemical factories. *Curr. Opin. Biotechnol*. **19**, 468-74. doi:10.1016/j.copbio.2008.07.009
9. Henry,C.S., Broadbelt,L.J., and Hatzimanikatis,V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol Bioeng*. **106**, 462-73. PMID: 20091733

10. Faulon,J.L. and Carbonell,P, (2010) Reaction Network Generation, In *Handbook of Chemoinformatics Algorithms*. Chapman & Hall/CRC Series in Mathematical & Computational Biology. www.crcpress.com/product/isbn/9781420082920

11. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-31. doi:10.1093/bioinformatics/btg015

12. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al*. (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **28**, 935-42. doi:10.1038/nbt1210-1308c

13. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28**, 27-30. doi:10.1093/nar/28.1.27

14. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al*. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **39**, D38-51. PMID: 21097890

15. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., Mcnaught,A., Alcántara,P., Darsow,M., Guedj,M., and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic Acids Res.* **36**, D344-50. doi:10.1002/0471250953.bi1409s26

16. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **28**, 31. doi:10.1021/ci00057a005

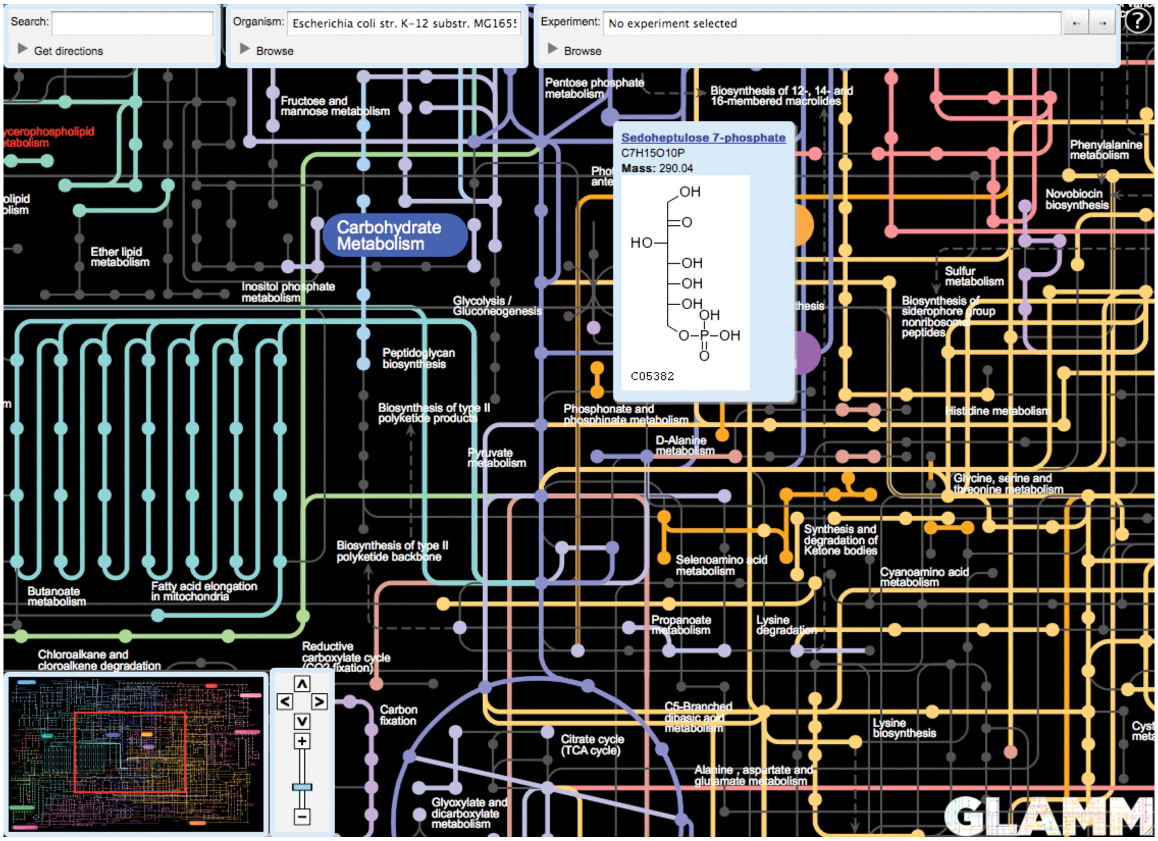
17. Stein,S.E., Heller,S.R., Tchekhovskoi,D. (2003) An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. In *Proceedings of the 2003 International Chemical Information Conference (Nimes), Infonortics*: 131-43. http://www.iupac.org/inchi/

18. Webb,E.C. (1992) Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Published for the *International Union of Biochemistry and Molecular Biology by Academic Press*. ISBN 0-12-227164-5. http://www.chem.qmul.ac.uk/iubmb/enzyme/

19. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R., and White,O. (2007) *Nucleic Acids Res*. **35**, D260-4. PMID: 17151080

20. Henry,C.S., DeJongh,M., Best,A.A., Frybarger,P.M., Linsay,B., and Stevens,R.L. (2010) High-throughput generation, optimization, and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977-82. doi:10.1038/nbt.1672

Figures



Search:

Organism: Escherichia coli str. K-12 substr. MG165

Experiment: No experiment selected

Sedoheptulose 7-phosphate

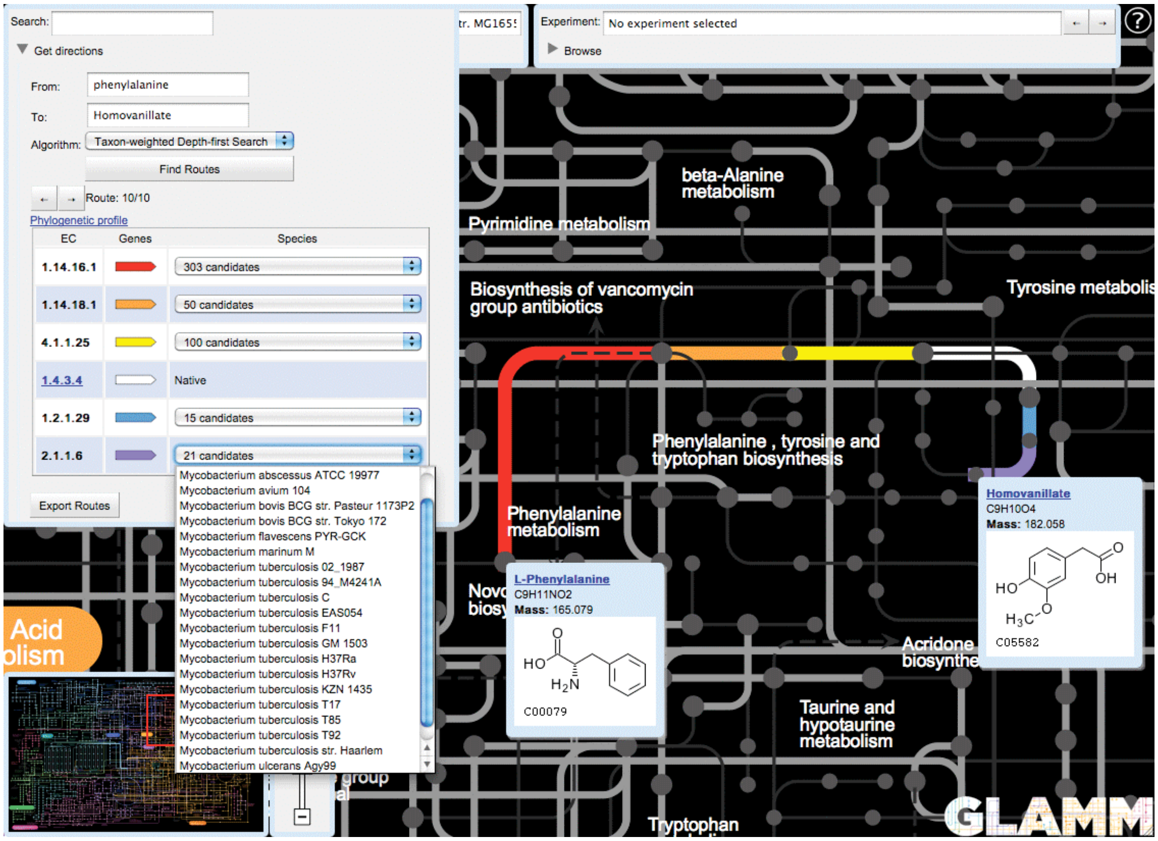
C7H15O10P

Mass: 290.04

C05382

Get directions
Browse
Browse
Fructose and mannose metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Phenylalanine metabolism
Novobiocin biosynthesis
Carbohydrate Metabolism
Ether lipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
Sulfur metabolism
Biosynthesis of siderophore group nonribosomal peptides
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Histidine metabolism
Pyruvate metabolism
D-Alanine metabolism
Glycine, serine and threonine metabolism
Synthesis and degradation of Ketone bodies
Biosynthesis of type II polyketide backbone
Selenoamino acid metabolism
Cyanoamino acid metabolism
Butanoate metabolism
Fatty acid elongation in mitochondria
Propanoate metabolism
Lysine degradation
Chloroalkane and cloroalkene degradation
Reductive carboxylate cycle (CO2 fixation)
Carbon fixation
C5-Branched dibasic acid metabolism
Lysine biosynthesis
Citrate cycle (TCA cycle)
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
OH
O
HO
OH
OH
OH
OH
O
P
OH
O
GLAMM

Figure 1. **Metabolic Reconstruction** of *E. coli K12 substr. MG1655* with metabolite information for Sedoheptulose 7-phosphate in a popup window. Reactions with genes identified in the reconstruction are shown in color, missing reactions in gray.



Search:

▼ Get directions

From: phenylalanine

To: Homovanillate

Algorithm: Taxon-weighted Depth-first Search

Route: 10/10

Phylogenetic profile

EC	Genes	Species
1.14.16.1	🟥	303 candidates
1.14.18.1	🟧	50 candidates
4.1.1.25	🟨	100 candidates
1.4.3.4	⬜	Native
1.2.1.29	🟦	15 candidates
2.1.1.6	🟪	21 candidates

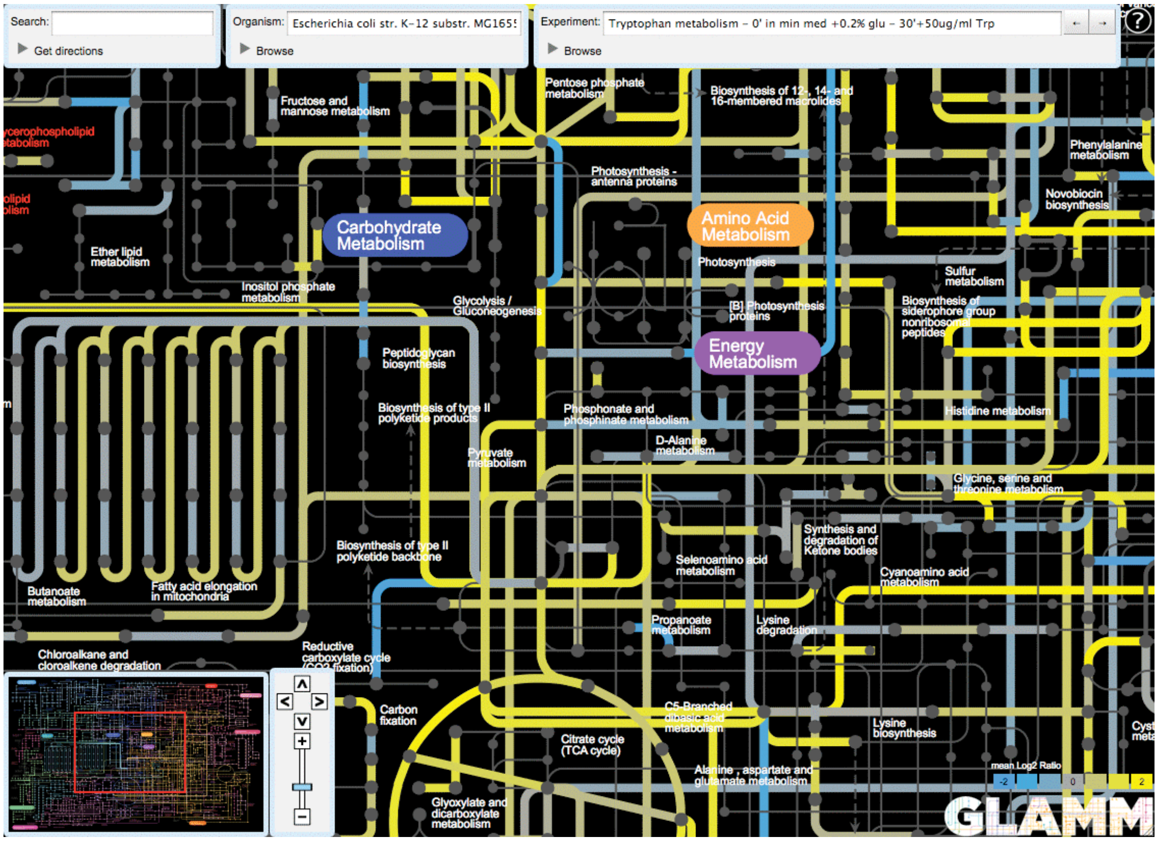
Mycobacterium abscessus ATCC 19977
Mycobacterium avium 104
Mycobacterium bovis BCG str. Pasteur 1173P2
Mycobacterium bovis BCG str. Tokyo 172
Mycobacterium flavescens PYR-GCK
Mycobacterium marinum M
Mycobacterium tuberculosis 02_1987
Mycobacterium tuberculosis 94_M4241A
Mycobacterium tuberculosis C
Mycobacterium tuberculosis EAS054
Mycobacterium tuberculosis F11
Mycobacterium tuberculosis GM 1503
Mycobacterium tuberculosis H37Ra
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis KZN 1435
Mycobacterium tuberculosis T17
Mycobacterium tuberculosis T85
Mycobacterium tuberculosis T92
Mycobacterium tuberculosis str. Haarlem
Mycobacterium ulcerans Agy99

Experiment: No experiment selected

L-Phenylalanine
C9H11NO2
Mass: 165.079
C00079

Homovanillate
C9H10O4
Mass: 182.058
C05582

Figure 2. Route Finding and Retrosynthesis "Getting directions" between the metabolites L-Phenylalanine and Homovanillate using *E. coli K12 substr. MG1655* as the host organism. Both endogenous (white) and exogenous reactions (colored) are shown, including the species names for the source of candidate genes for the transgenic steps in the pathway.



Search:

Organism: Escherichia coli str. K-12 substr. MG165!

Experiment: Tryptophan metabolism – 0' in min med +0.2% glu – 30'+50ug/ml Trp

Figure 3. Experimental Data Visualization. Overlay of expression data collected during a metabolism experiment on *E. coli K12 substr. MG1655*. The reactions corresponding to upregulated genes are shown in yellow, reactions corresponding to downregulated genes are shown in blue.